Simple Linear Regression

Yesoda Bhargava
MPH, University of York, UK

February 12, 2020

Background

- To analyse two continuous quantitative variables Pearson correlation coefficient (R) is used and scatter plot to depict the information graphically.
- However, using the correlation coefficient (R) we cannot quantify the change in relative estimate of one variable as another variable changes.
- Note that correlation coefficient does not entail the labeling of one variable as explanatory/independent and response/dependent.
- Thus, to quantify the change mentioned above, regression is used. Regression can be of two types: **Simple Linear Regression** and **Multiple Linear Regression**. It is also non-linear type in which case it is called as **Curvilinear Regression**.

Dataset used

- *birthwt* data from MASS package in R.
- Explanatory variable: *lwt* - mother's weight in pounds at last menstrual period.
- Response variable: *bwt* - birth weight in grams.
- To load dataset in R:
  library(MASS)
  head(birthwt)

```
> head(birthwt)
   low age lwt race smoke ptl ht ui ftv  bwt
85   0  19 182    2     0   0  0  1   0 2523
86   0  33 155    3     0   0  0  0   3 2551
87   0  20 105    1     1   0  0  0   1 2557
88   0  21 108    1     1   0  0  1   2 2594
89   0  18 107    1     1   0  0  1   0 2600
91   0  21 124    3     0   0  0  0   0 2622
```

Figure 1: Snapshot of the input data

Correlation

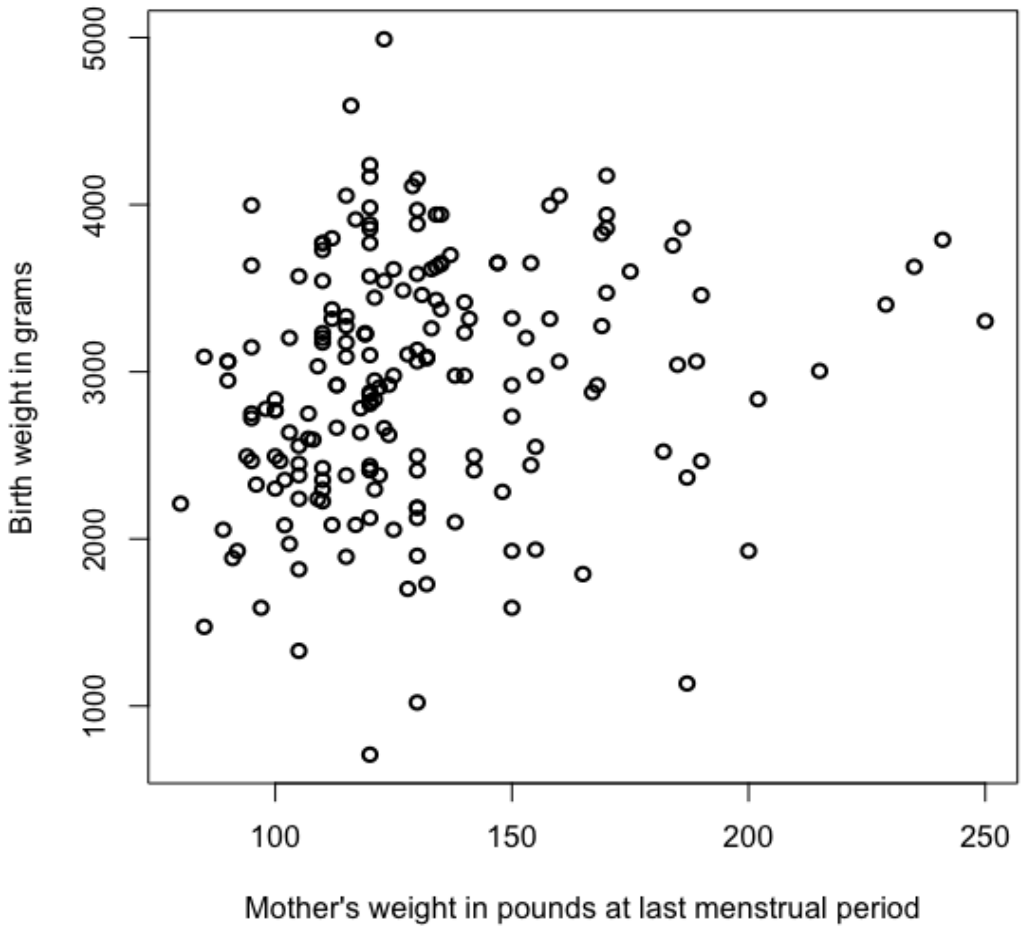

Figure 2: Baby birth weight vs. Mother's weight at last menstrual period. Correlation coefficient is $= 0.185$

Introduction to Simple Linear Regression

- Looking at Fig. 2, we cannot estimate how much change in birth weight (bwt) occurs if mother's weight (lwt) changes. For that we use Regression. Note that, here the explanatory variable is mother's weight (lwt) and the response variable is birth weight of child (bwt).
- Since, we have only one explanatory variable, the regression is called as Simple Linear Regression; "Simple" indicates single explanatory variable.
- The Regression model/equation assumes that there exists a theoretical linear model in the population which can explain the variability in Y (response variable) using X (explanatory variable). In the current situation Y is represented by baby birth weight (bwt) and X by mother's weight (lwt).

- The population equation model can be given by:

$$\mu_{y|x} = \beta_0 + \beta_1 x \tag{1}$$

where $\mu_{y|x}$ denotes the average value of y given x, β_0 is the constant and β_1 is the coefficient of x. The simple linear equation can be assumed to be a line that is fit in the 2-D space between explanatory and response variable. The general equation of a line is

$$y = mx + c \tag{2}$$

where m is the slope and c is the intercept. Thus, in the simple linear regression equation β_0 corresponds to c and β_1 corresponds to m.

- The population equation (1) is estimated by fitting the model of the form

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{3}$$

where ε known as the error or residual is the distance of a particular outcome y from the population regression line (1).

- After fitting the line we obtain the least-squares regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{4}$$

Least squares regression

- Many possible lines may explain the relationship between y and x as shown in Fig. 2.

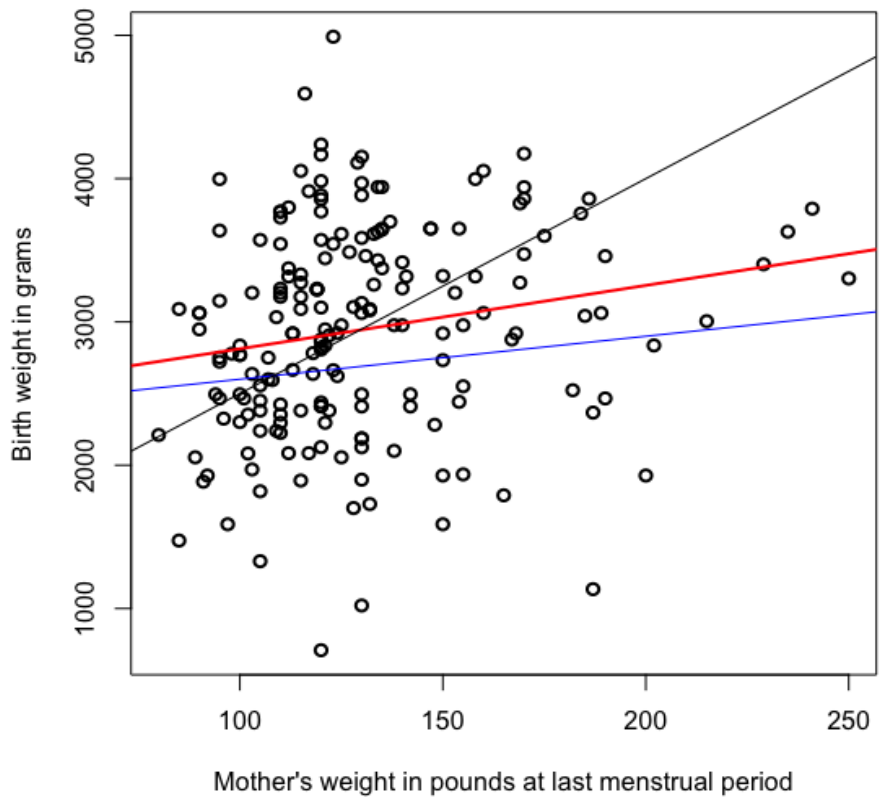

Figure 3: Multiple Regression line in scatter plot

Least squares regression

- We need to identify a line which gives the least error in the predicted and actual values of y. $\varepsilon_i = y_i - \hat{y}_i$. The goal is to minimise the squared sum of these errors over all data points, say n, i.e.

$$\sum_{i=1}^{n} \varepsilon_i = (y_i - \hat{y}_i)^2 \tag{5}$$

Equation (5) can also be written as

$$\sum_{i=1}^{n} \varepsilon_i = (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \tag{6}$$

The sum is minimised using partial differentiation and the coefficient values of β_0 and β_1 are obtained. This is why the obtained line is called as the Least Squares Regression Line.

Least squares regression

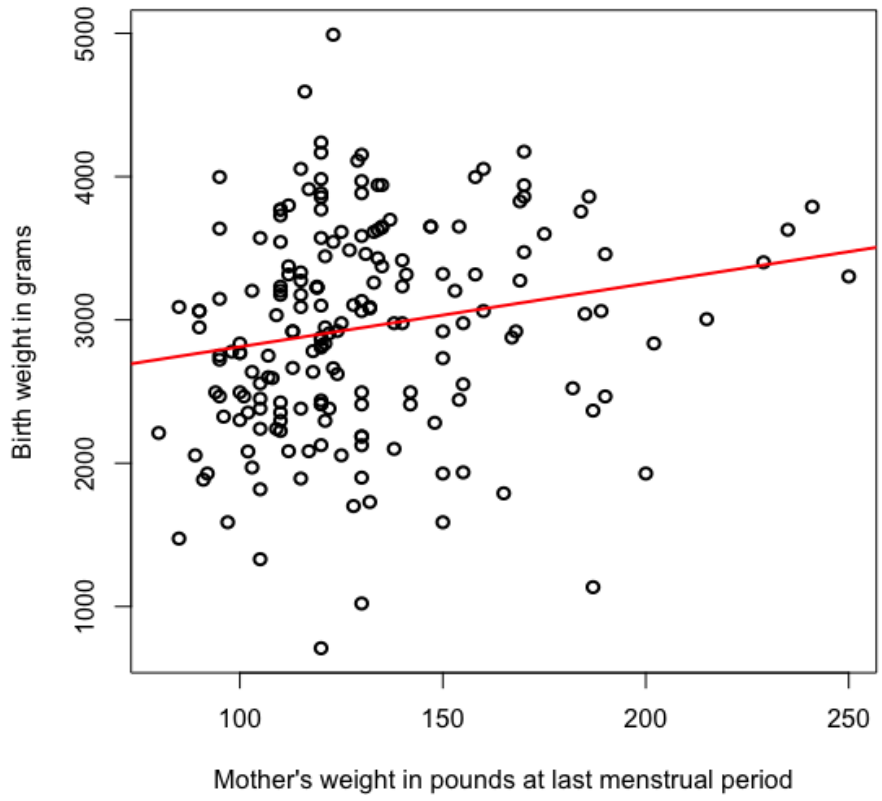

Figure 4: Least Squares Regression line in scatter plot

Estimated Regression Line

- The obtained least squares regression line is:

$$\hat{y} = 2369.624 + 4.429x \tag{7}$$

  and it has a property that sum and mean of residuals is approximately zero.
- Equation (7) can be used to estimate the **mean** y for a sub-population having a particular value of x or **individual value of** y for the single value of x. The error in former case is lesser because it is the mean outcome whereas to predict the single outcome, the prediction error is more.
- The estimate of y is the same in both the cases only error or the confidence interval of the estimate differs.

Interpretation of the coefficients

- The obtained least squares regression line is:
  $\hat{y} = 2369.624 + 4.429x$
  Note that the above equation estimates population regression equation (1). Thus, the output is the average output at a particular value of x. Let's assume initial value of x ix x_0, so the mean outcome is: $\hat{y}_1 = 2369.624 + 4.429x_0$.
  Now we change the value of x_0 and increase it by one unit, obtaining $x_0 + 1$. The obtained mean output is
  $\hat{y}_2 = 2369.624 + 4.429(x_0 + 1)$.
  Subtracting $\hat{y}_1$ from $\hat{y}_2$, gives us a value of 4.429 which is nothing but the coefficient for x, β_1.
- β_1 is thus, the change in mean value of outcome y for one unit increase in x.
- Now put, $x = 0$ in equation (7), we obtain $\hat{y} = 2369.624$. Thus, β_0 is the mean value of y at x=0, i.e., when the value of explanatory variable's value is zero. Note that in this case, lwt=0 makes no sense, hence the constant does not convey any information.

Assumptions of Linear Regression

To check the assumptions of the simple linear regression we look at the residual plots. Assumptions that need to be checked are

- Independence of residuals. (Study design, internal validity of the study)
- Residuals are normally distributed. (Check using Normal Q-Q Plot).
- Uniformity of variance / homoscedasticity: Homoscedasticity means that the standard deviation of the outcomes y is constant across all values of x. If the range of the magnitude of the residuals either increases or decreases as $\hat{y}$ becomes larger producing a fan-shaped scatter - it implies that the standard deviation does not take the same value for all values of x. If this happens, the assumption of homoscedasticity is violated.

Residual plots

Residuals are plots of residuals versus fitted/predicted $\hat{y}$ values and can serve 3 purposes:

- Detect outlying observations in the sample. Like sample means ($\bar{X}$) and correlation coefficient, method of least squares can be very sensitive to outliers in the data especially if they correspond to relatively large or small values of x. Care must be taken not to throw away unusual data points that are in fact valid; these observations might be the most interesting ones in the data.
- Suggest failure of assumption of homoscedasticity. A failure would indicate that simple linear regression is not the appropriate technique for modelling the relationship between x and y. (Note that it can be difficult to evaluate this and other assumptions based on residual plot if the number of data points is small).

Residual plots

To check the assumptions of the simple linear regression we look at the residual plots. Assumptions that need to be checked are:

- If residuals do not exhibit a random scatter but instead follow a distinct trend : ε_i increases as $\hat{y}_i$ increases, this suggests that the true relationship might not be linear. In this case a transformation of one of both variables may be appropriate.

Normality of residuals assumption

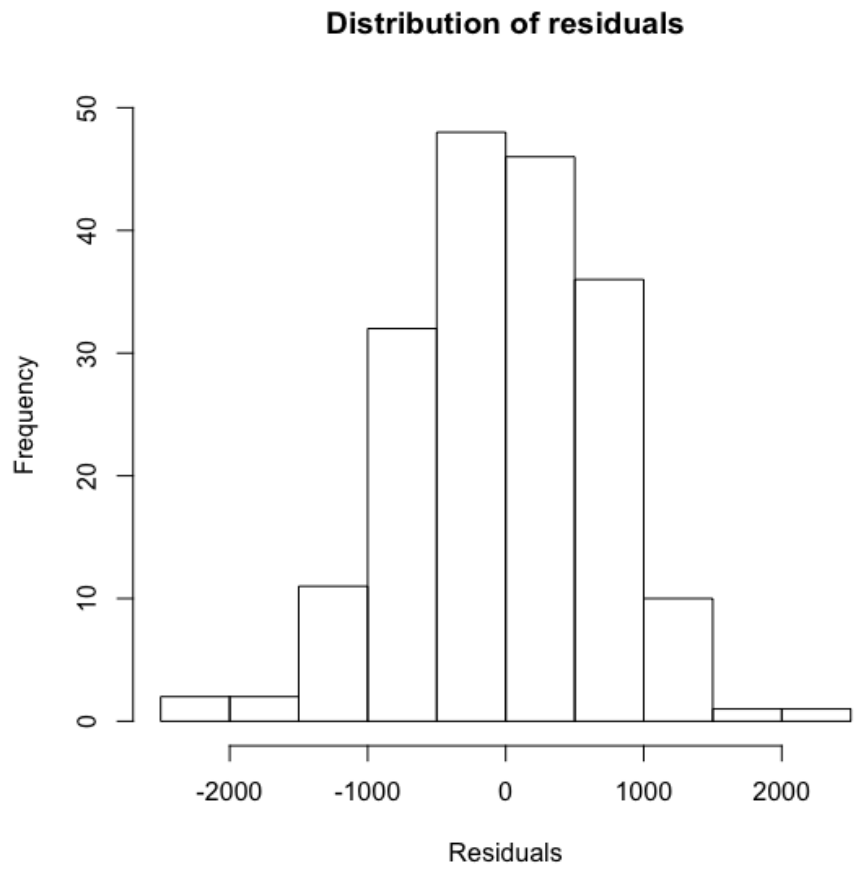

Figure 5: Distribution of residuals. Checking for normality of histogram.

Contd.

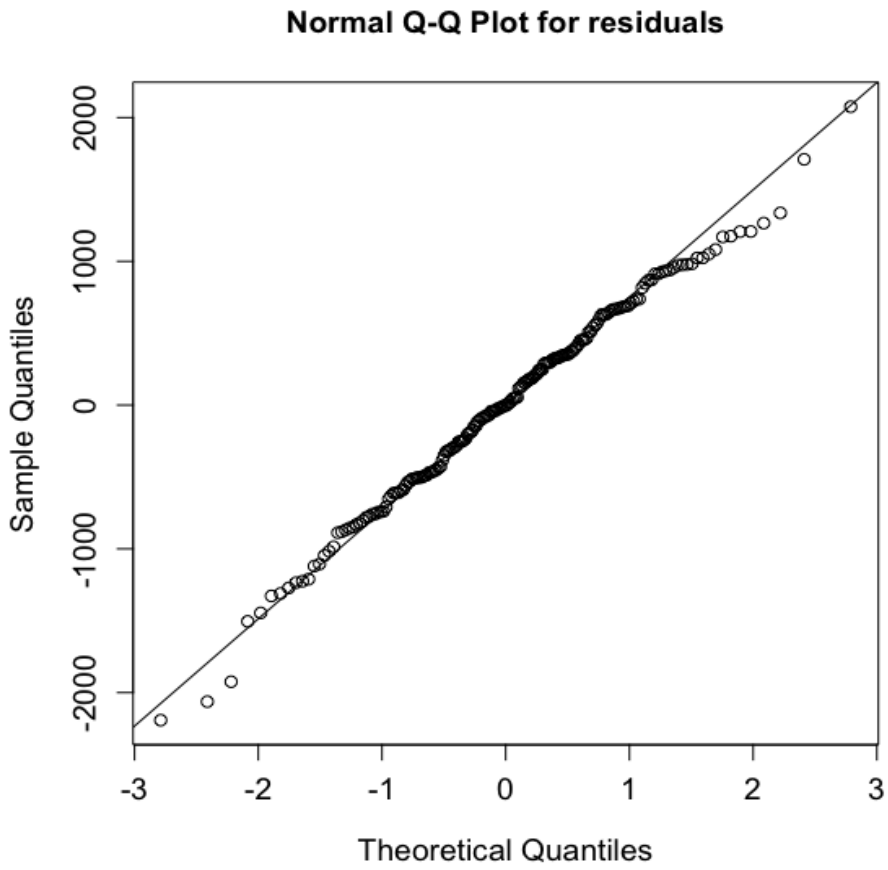

Figure 6: Checking normality of residuals. (Recommended method)

Uniformity of variance/Homoscedasticity test

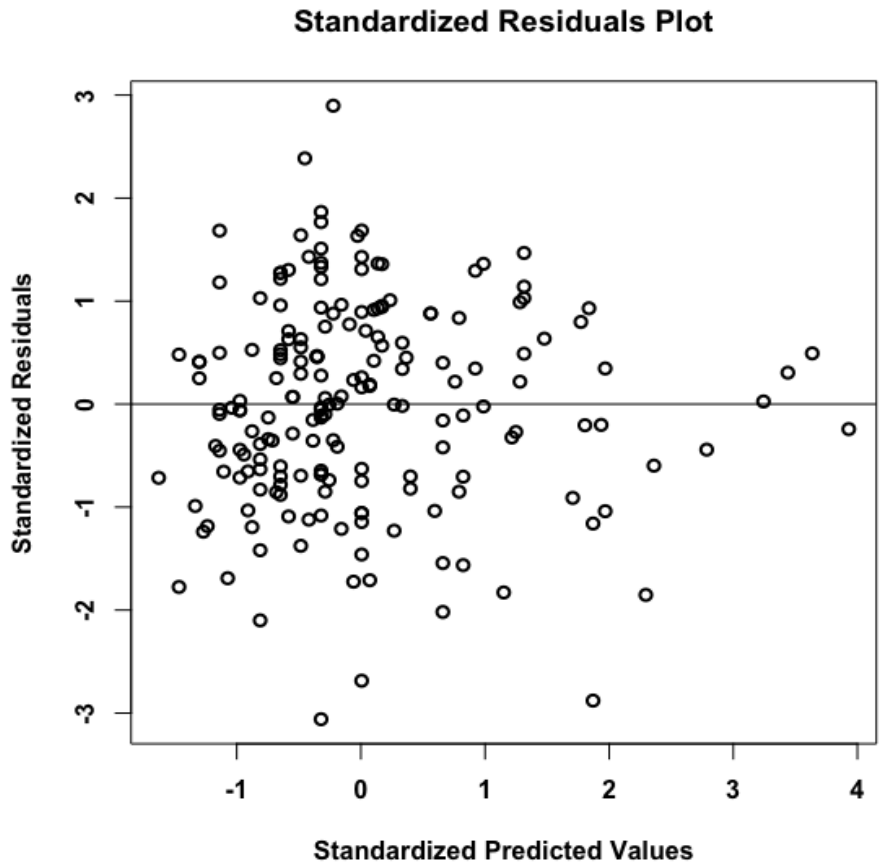

Figure 7: Standardized Residuals vs. Standardized Predicted Values

Practical implementation

Visit Youtube Channel.
Please write to yesodabhargava@gmail.com for any confusions, corrections and suggestions. Thank you.